Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

A Case Study on Sepsis Using PubMed and Deep Learning for Ontology Learning

Mercedes ARGUELLO CASTELEIRO[a], Diego MASEDA FERNANDEZ[b], George DEMETRIOU[a], Warren READ[a], Maria Jesus FERNANDEZ PRIETO[c], Julio DES DIZ[d], Goran NENADIC[a,e], John KEANE[a,c], and Robert STEVENS[a,1]
[a]*School of Computer Science, University of Manchester (UK)*
[b]*Midcheshire Hospital Foundation Trust, NHS England (UK)*
[c]*Salford Languages, University of Salford (UK)*
[d]*Hospital do Salnés de Villagarcia, SERGAS (Spain)*
[e]*Manchester Institute of Biotechnology, University of Manchester (UK)*

Abstract. We investigate the application of distributional semantics models for facilitating unsupervised extraction of biomedical terms from unannotated corpora. Term extraction is used as the first step of an ontology learning process that aims to (semi-)automatic annotation of biomedical concepts and relations from more than 300K PubMed titles and abstracts. We experimented with both traditional distributional semantics methods such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) as well as the neural language models CBOW and Skip-gram from Deep Learning. The evaluation conducted concentrates on sepsis, a major life-threatening condition, and shows that Deep Learning models outperform LSA and LDA with much higher precision.

Keywords. Ontology Learning, Deep Learning, PubMed, OWL, SPARQL

1. Introduction

Sepsis is defined as a life-threatening organ dysfunction caused by a dysregulated host response to a new infection [1]. *Severe sepsis* occurs when complicated with organ dysfunction, and may progress to *septic shock* [1]. *Severe sepsis* and *septic shock* affect millions of people worldwide, and their incidence is increasing [2]. According to NHS England: "*sepsis now claims more lives than lung cancer, the second biggest cause of death after cardiovascular disease*" [3]. Severe sepsis is regarded as "*a common, deadly, and costly complication in cancer patients*" [4].

The Surviving Sepsis Campaign (SSC), established in 2002, updates consensus definitions and evidence-based guidelines for management of severe sepsis and septic shock. Evidence from the literature improves understanding of sepsis, leading to better patient care. However, the size and rate of growth of PubMed – the largest biomedical resource – is a challenge for term extraction and knowledge representation.

Ontology learning from text aims to "*turn facts and patterns from an ever growing body of information into shareable high-level constructs*" [5]. This paper investigates how the traditional distributional semantics methods, Latent Semantic Analysis (LSA)

[1] Corresponding author, E-mail: Robert.Stevens@manchester.ac.uk

[6] and Latent Dirichlet Allocation (LDA) [7], as well as the neural language models CBOW (Continuous Bag-of-Words) and Skip-gram of Mikolov et al. [8] from Deep Learning can facilitate ontology learning from an unannotated biomedical corpus.

2. Methods

In distributional semantic models (DSMs), the semantics of terms within a domain are determined empirically [9]. LSA is spatially motivated, while LDA is a probabilistic method. LSA and LDA have high computational and storage cost associated with building or modifying the model. However, CBOW and Skip-gram from Deep Learning make it feasible to obtain word embeddings (i.e. distributed word representations) from corpora of billions of words. Using similarity measures (e.g. cosine value for CBOW and Skip-gram) we can obtain the n top-ranked candidate terms for a query term (e.g. sepsis). Hence we can quantify empirically how closely related are two terms from a DSM.

This study takes advantage of the UMLS (Unified Medical Language System) Metathesaurus from the U.S. National Library of Medicine that contains more than three million biomedical concepts. Each UMLS concept has a unique identifier (a.k.a. CUI). UMLS Metathesaurus concepts are grouped into UMLS Semantic Types, where semantic types can also be merged into semantic groups [10].

We use OWL-DL [11] to formally represent concept names, concept expressions, and terminological axioms. We started by creating programmatically a small ontology that contains mostly the UMLS Semantic Types and Groups. This ontology together with lemon (Lexicon Model for Ontologies) [12] is key to the move from *candidate terms* from DSMs to UMLS Metathesaurus concepts with known synonyms and relationships. In this study, the concept *Lexicon* from lemon represents a vocabulary for a DSM, while the concept *Lexical entry* represents a single term (one or more words) in the vocabulary/lexicon of the DSM. We also link the concept *Lexical sense* from lemon with the UMLS Metathesaurus concept, an OWL class we created that can have as a subclass any of the more than 3 million UMLS Metathesaurus concepts.

A vocabulary/lexicon from DSMs contains lexical entries that are: concepts, phraseological expressions (typically combination of concepts), or spurious terms without a true biomedical meaning. UMLS MetaMap [13] can indicate which terms from the lexicon are UMLS Metathesaurus concepts and their UMLS Semantic Type(s).

Experimental setup – We downloaded the MEDLINE/PubMed baseline files for 2015 and also the update files up to 8th June 2016. Applying the PubMed Systematic Reviews filter [14], a subset of 301,202 PubMed publications (title and abstract) with date of publication from 2000 to 2016 was obtained. We performed two experiments using LSA, LDA, CBOW and Skip-gram: *Experiment I* – the pre-processing performed preserves capitalisation and numbers in the text; and *Experiment II* – after the pre-processing that preserves capitalisation and numbers in the text, Part-Of-Speech (POS) tagging and chunking was performed. Chunking labels segments of a sentence with syntactic constituents – e.g. noun phrase (NP) and verb phrase (VP). For the experiments reported here, we used gensim [15] and word2vec [16].

Domain expert evaluation – three medical consultants (rater A, B, and C) assessed the relevance of the terms in pairs (query term and candidate term) using a Likert-type (categorical) scale taken from [17]. According to this scale, a candidate term can be: not at all relevant (marked as 0); a little relevant (marked as 1); quite a bit

relevant (marked as 2); and very much relevant (marked as 3). Simple guidelines were given to the domain experts. They consist of: a) the Likert-type (categorical) scale; and b) a few examples illustrating pairs of query term-candidate term annotated with different scores. The inter-annotator agreement is calculated using two well-known measures: weighted Cohen's Kappa and Fleiss. Using few raters and simplistic annotation guidelines, low inter-annotator agreement and some difficulties are expected before arriving at consistent figures [18].

Task-based evaluation – the automatically extracted ontology – called here PubMed Ontology LEarning Ontology for Sepsis (POLEOS) – refers to the OWL-DL ontology built programmatically out of the n top-ranked candidate terms obtained from each model (i.e. LSA, LDA, CBOW, and Skip-gram) in *Experiment I* and *II*. POLEOS reuses the UMLS Semantic Types and Groups formalised in OWL. Two annotation properties were introduced in POLEOS to capture: a) the relevance of a candidate term, which comes from a voting system that considers equally the assessment made by each human rater per candidate term; and b) to what extent the candidate term was recognised by UMLS MetaMap. Using conventional measures in information retrieval [5], the suitability of the different models to find relevant terms (concepts or phraseological expressions) for sepsis is assessed. The numbers needed to calculate precision are obtained by executing SPARQL [19] queries over POLEOS. We can also query POLEOS with SPARQL to determine: the number of unique UMLS concepts needed to provide the biomedical meaning for the candidate terms; and how many candidate terms can be assigned to each UMLS Semantic type or group based on the *Lexical sense* assigned.

3. Results

Computational resources and execution times – The DSMs are generated using a Supermicro with 256GB RAM and two CPUs Intel Xeon E5-2630 v4 at 2.20GHz. The time to obtain a DSM goes from less than 1 hour for CBOW (the quickest) to more than 23 hours for LDA (the slowest). Using a MacBook Pro Retina (2.8 GHz Intel Core i7 with 16GB RAM) and Jena ARQ [20] as the SPARQL query engine, the mean time for executing a SPARQL query over POLEOS three times was less than 2 seconds.

Distributional Semantics: LSA, LDA, CBOW, and Skip-gram –The query term was "sepsis" for all models. For LDA in *Experiment II*, two topics for sepsis were identified, and thus, LDA in *Experiment II* has almost double the candidate terms than the other models.

Domain expert evaluation – Rater C did not score any candidate term with 0 (not at all relevant); while rater A scored 52 of the candidate terms with 0. Based on this finding, we decided to merge the four scores into two categories: *more relevant* with score 3 or 2; and *less relevant* with score 1 or 0. These two categories allow arrival at consistent figures, where the weighted Cohen's Kappa between two consultants is: 0.53 for rater A and B; 0.61 for rater A and C; and 0.55 for rater B and C. Fleiss Kappa is 0.56 for the three raters. In Figure 1, each column corresponds to a model with its name at the top. A column has up to three colours: dark grey for *most relevant* (score 3) candidate terms agreed by all three raters, grey for *more relevant* (score 3 or 2) agreed by at least two raters, and white *less relevant* (score 1 or 0) candidate terms agreed by at least two raters.

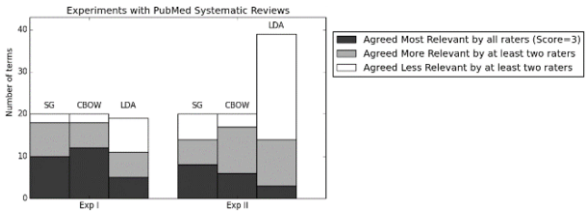

Figure 1. Agreed assessment by medical consultants for candidate terms, where SG = Skip-gram

Task-based evaluation – POLEOS contains a total of 3276 axioms, and its Description Logic expressivity is ALEHI(D). From the two experiments and the different models, a total of 157 candidate terms related to sepsis were obtained. The same candidate term may appear in more than one model; 130 candidate terms are unique: 71 are UMLS concepts, 50 are phraseological expressions (typically a combination of UMLS concepts), and 9 were unrecognised by UMLS MetaMap. Only 7 of the 9 unrecognised are truly spurious terms (i.e. terms that do not have a true biomedical meaning); the other two (i.e. suPAR and IgGAM) are acronyms that could be mapped to UMLS concepts. A total of 169 unique UMLS concepts are represented as subclasses of the concept *Lexical sense* from lemon. In other words, 169 UMLS concepts are needed to provide the biomedical meaning of the 157 candidate terms obtained. Three UMLS Semantic groups stand out with more candidate terms (concepts or phraseological expressions): *Disorders* (DISO) with 61 candidate terms, *Concepts & Ideas* (CONC) with 47 candidate terms, and *Living Beings* (LIVB) with 24 candidate terms. Table 1 shows the precision for each model in *Experiment I* and *II*.

Table 1. Task-based precision calculated as tp/(tp+fp) per model and experiment, where tp stands for the number of true positives agreed by at least two raters and fp stands for the number of false positives agreed by at least two raters.

Model	Experiment I	Experiment II
LSA	-	26%
LDA	58%	36%
CBOW	**90%**	**85%**
Skip-gram	**90%**	70%

4. Discussion and Conclusion

Healthcare data is multi-source, high volume and multi-modal. Identifying patterns in such data requires scalability while accommodating structured and unstructured data. Unlike conventional datasets, healthcare data is often incomplete and noisy; in turn, unlike standard analytics, neural language models process raw natural language data to associate terms with vectors of real valued features and place semantically related terms close together in the vector space [21]. The learned "*high level*" semantic features of the word embeddings are usually not explicitly present in input such as biomedical literature.

Hu et al. [22] recently reported improvements in the word embeddings generated by introducing POS tagging information into a neural network similar to CBOW. However, from the precision obtain for Experiment II (see Table 1) it is difficult to derive a real benefit from the chunking (VP and NP) performed.

Although several parameter settings were tried, it is plausible that other options may have improved the precision for LSA and LDA. However, a higher performance for CBOW and Skip-gram when compared with traditional DSMs such as LDA is aligned with the results reported in the literature for other studies, e.g. [23].

Using sepsis as the query term, we have shown how to anchor plain text candidate terms from DSMs into shareable high-level constructs, i.e. UMLS concepts represented in OWL that can be easily queried with SPARQL based on Semantic types and Semantic Groups. The evaluation performed indicates high precision for the neural language models, which in turn hints at the plausible acquisition of relevant medical concepts for sepsis. Hence, this paper illustrates how CBOW and Skip-gram can be used to aid ontology learning tasks for sepsis, a major healthcare problem, using unannotated PubMed Systematic Reviews citations (titles and abstracts) as a corpus.

Acknowledgements

This work was supported by grant 603288 from the European Union Seventh Framework Programme (FP7/2007-2013) for sysVASC project.

References

[1] Singer, M., et al. 2016. The third international consensus definitions for sepsis and septic shock (sepsis-3). Jama 315 (2016), 801-810.
[2] Levy, M.M., et al. Surviving Sepsis Campaign: association between performance metrics and outcomes in a 7.5-year study. Intensive care medicine 40 (2014), 1623-1633.
[3] NHS Report, https://www.england.nhs.uk/wp-content/uploads/2015/08/Sepsis-Action-Plan-23.12.15-v1.pdf Accessed Feb 2017.
[4] Williams, M.D., et al.,. Hospitalized cancer patients with severe sepsis: analysis of incidence, mortality, and associated costs of care. Critical Care, 8 (2004).
[5] Wong, W., Liu, W. and Bennamoun, M. Ontology learning from text: A look back and into the future. ACM Computing Surveys (CSUR) 44 (2012).
[6] Landauer, T.K. and Dumais, S.T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review 104 (1997).
[7] Blei, D.M., Ng, A.Y. and Jordan, M.I. Latent dirichlet allocation. Journal of machine Learning research 3 (2003), 993-1022.
[8] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. arXiv pre-print arXiv:1301.3781 (2013).
[9] Cohen, T. and Widdows, D. Empirical distributional semantics: methods and biomedical applications. Journal of biomedical informatics 42 (2009), 390-405.
[10] UMLS Semantic Network, https://semanticnetwork.nlm.nih.gov Accessed Feb 2017.
[11] OWL 2 Web Ontology Language, https://www.w3.org/TR/owl2-overview/ Accessed Feb 2017.
[12] lemon - The Lexicon Model for Ontologies, http://lemon-model.net Accessed Feb 2017.
[13] MetaMap, https://metamap.nlm.nih.gov Accessed Feb 2017.
[14] PubMed SB, https://www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_strategy.html Accessed Feb 2017.
[15] Gensim, https://radimrehurek.com/gensim/ Accessed Feb 2017.
[16] word2vec, https://code.google.com/archive/p/word2vec/ Accessed Feb 2017.
[17] Aaronson, N.K. Quality of life assessment in clinical trials: methodologic issues. Controlled Clinical Trials 10 (1989), 195-208.
[18] Biemann, C. Ontology learning from text: A survey of methods. In LDV forum 20 (2005), 75-93.
[19] SPARQL 1.1 query language, https://www.w3.org/TR/sparql11-query/ Accessed Feb 2017.
[20] Jena ARQ, http://jena.sourceforge.net/ARQ/ Accessed Feb 2017.
[21] LeCun, Y., Bengio, Y. and Hinton, G. Deep learning. Nature, 521 (2015), 436-444.
[22] Hu, B., Tang, B., Chen, Q. and Kang, L. A novel word embedding learning model using the dissociation between nouns and verbs. Neurocomputing, 171 (2016), 1108-1117.
[23] Liu, Y., Liu, Z., Chua, T.S. and Sun, M. Topical Word Embeddings. In AAAI (2015), 2418-2424.